

Enforcing Calibration in Multi-Output Probabilistic Regression with Pre-rank Regularization

Naomi Desobry^{1*}, Elnura Zhalieva^{2*}, Souhaib Ben Taieb^{1,2}

¹Department of Computer Science, University of Mons

²Department of Statistics and Data Science, Mohamed Bin Zayed University of Artificial Intelligence
naomi.desobry@umons.ac.be, elnura.zhalieva@mbzuai.ac.ae, souhaib.bentaieb@mbzuai.ac.ae

Abstract

Probabilistic models must be well calibrated to support reliable decision-making. While calibration in single-output regression is well studied, defining and achieving multivariate calibration in multi-output regression remains considerably more challenging. The existing literature on multivariate calibration primarily focuses on diagnostic tools based on pre-rank functions, which are projections that reduce multivariate prediction-observation pairs to univariate summaries to detect specific types of miscalibration. In this work, we go beyond diagnostics and introduce a general regularization framework to enforce multivariate calibration during training for arbitrary pre-rank functions. This framework encompasses existing approaches such as highest density region calibration and copula calibration. Our method enforces calibration by penalizing deviations of the projected probability integral transforms (PITs) from the uniform distribution, and can be added as a regularization term to the loss function of any probabilistic predictor. Specifically, we propose a regularization loss that jointly enforces both marginal and multivariate pre-rank calibration. We also introduce a new PCA-based pre-rank that captures calibration along directions of maximal variance in the predictive distribution, while also enabling dimensionality reduction. Across 18 real-world multi-output regression datasets, we show that unregularized models are consistently miscalibrated, and that our methods significantly improve calibration across all pre-rank functions without sacrificing predictive accuracy.

Introduction

Probabilistic models output full predictive distributions rather than point estimates, enabling principled uncertainty-aware decision-making in domains such as meteorology, finance, and medical diagnosis (Murphy and Winkler 1984; Krzysztofowicz and Evans 2008; de Lima Silva et al. 2020; Önköl and Muradoğlu 1994; Gulshan et al. 2016; Guizilini et al. 2019). However, to be reliable, these predictions must be *calibrated*, that is, their predicted probabilities must align with the observed frequencies.

In single-output regression, calibration is well understood and can be evaluated using tools such as the Probability Integral Transform (PIT). Deviations from perfect calibration, referred to as miscalibration, can be corrected either during

training via regularization techniques based on scoring rule decompositions (Wilks 2018; Wessel et al. 2025), or post hoc using recalibration methods such as isotonic regression (Kuleshov, Fenner, and Ermon 2018) or kernel-based adjustments (Dheur and Taieb 2023). Throughout this work, we refer to calibration in single-output regression as univariate calibration.

Multivariate calibration, by contrast, concerns the calibration of a multivariate target and is considerably more difficult to evaluate and achieve. When producing probabilistic predictions for such targets, correctly specifying the marginal distributions is not sufficient; the predictions must also accurately capture the dependencies and joint structure across target dimensions. Although several tools have been proposed to assess specific aspects of multivariate calibration (Chung, Char, and Schneider 2024; Ziegel and Gneiting 2013), defining general-purpose, interpretable, and effective calibration methods for the multivariate setting remains an open challenge.

One approach to evaluating multivariate calibration involves the use of pre-rank functions, which are scalar summaries of prediction-observation pairs that extend univariate rank-based diagnostics to the multivariate setting (Allen, Ziegel, and Ginsbourger 2023). Each pre-rank targets a specific aspect of miscalibration, such as marginal calibration or discrepancies in summary statistics like location, scale, or dependence structure. By projecting complex multivariate predictions onto interpretable scalar quantities, pre-rank functions provide a flexible and general framework for assessing different dimensions of probabilistic calibration.

In this work, we go beyond diagnostic tools and propose a method to directly enforce *multivariate calibration* by incorporating a regularization term into the training loss. This term penalizes miscalibration with respect to a collection of pre-rank functions. We further introduce a novel pre-rank based on Principal Component Analysis (PCA), which projects prediction-observation pairs onto directions of maximal variance in the predictive distribution, thereby capturing calibration along statistically meaningful directions. Additionally, we propose a regularization loss that jointly enforces both marginal and multivariate pre-rank calibration. When combined with the PCA-based pre-rank, our approach also enables dimensionality reduction and improves computational efficiency. Empirically, our method consistently

*These authors contributed equally.
Preprint. Under review.

improves calibration across all pre-rank functions without compromising predictive accuracy. We make the following main contributions:

- We conduct a large-scale empirical study on 18 real-world multi-output regression datasets to evaluate the probabilistic calibration of unregularized models across a diverse set of pre-rank functions.
- We propose a general regularization framework that can be integrated into the training of any probabilistic predictor to enforce multivariate calibration with respect to user-specified pre-rank functions. Our approach also includes a joint regularization loss that enforces both marginal and multivariate calibration. When combined with our PCA-based pre-rank, the method detects calibration along the top principal components of the predictive covariance while also serving as a dimensionality reduction technique.
- We validate our framework on 18 benchmark datasets and show that it consistently improves calibration across all pre-rank metrics without compromising predictive accuracy.

Background

We consider a multivariate regression setting where inputs $X \in \mathcal{X} \subseteq \mathbb{R}^L$ and targets $Y \in \mathcal{Y} \subseteq \mathbb{R}^D$ are jointly distributed. The target $Y = (Y_1, \dots, Y_D)$ has dimension $D \geq 1$. Our goal is to estimate the true conditional distribution $F_{Y|X}$ from a finite dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$.

To this end, we define a *probabilistic predictor* $F_\theta : \mathcal{X} \rightarrow \mathcal{F}$, where $\mathcal{F} \subseteq \mathcal{P}(\mathbb{R}^D)$ is a class of admissible probability distributions over $\mathbb{R}^D$, $\mathcal{P}(\mathbb{R}^D)$ denotes the space of all probability distributions on $\mathbb{R}^D$, and θ represents the model parameters. For any input $x \in \mathcal{X}$, the model outputs a predictive cumulative distribution function $\hat{F}_{Y|X=x} \in \mathcal{F}$, with corresponding density $\hat{f}_{Y|X=x}$. This distribution may be available in closed form (e.g., a multivariate Gaussian) or approximated via samples.

We learn the model parameters θ by minimizing a proper scoring rule over a training dataset, thereby encouraging the predictive distribution $F_\theta(x)$ to align with the true conditional distribution of Y given $X = x$. A scoring rule $S : \mathcal{F} \times \mathcal{Y} \rightarrow \mathbb{R}$ assigns a numerical score to each predictive distribution $\hat{F} \in \mathcal{F}$ and observed outcome $y \in \mathcal{Y}$. It is called *proper* if it is minimized in expectation when $\hat{F}$ equals the true distribution, and *strictly proper* if the minimizer is unique. Two widely used examples are the negative log-likelihood (NLL), $\text{NLL}(\hat{F}, y) = -\log \hat{f}(y)$, where $\hat{f}$ is the density of $\hat{F}$, and the energy score (ES),

$$\text{ES}(\hat{F}, y) = \mathbb{E}_{\hat{Y} \sim \hat{F}} \|\hat{Y} - y\| - \frac{1}{2} \mathbb{E}_{\hat{Y}, \hat{Y}' \sim \hat{F}} \|\hat{Y} - \hat{Y}'\|.$$

This estimation strategy, known as *optimum score estimation* (Gneiting and Raftery 2007), allows flexible learning of predictive distributions. However, model misspecification and limited data may lead to biased or miscalibrated models, and the choice of scoring rule can also affect the accuracy, robustness, and calibration of the resulting model.

Univariate calibration. To better understand calibration in the multivariate setting, we briefly recall probabilistic calibration in the univariate setting. Let $X \in \mathcal{X}$ and $Y \in \mathbb{R}$ be random variables with conditional distribution $F_{Y|X}$, and let $\hat{F}_{Y|X}$ be a probabilistic predictor.

Definition 1. $\hat{F}_{Y|X}$ is said to be PIT-calibrated if the probability integral transform (PIT),

$$Z = \hat{F}_{Y|X}(Y),$$

is uniformly distributed on $[0, 1]$, that is,

$$F_Z(\alpha) = \alpha \quad \text{for all } \alpha \in [0, 1]. \quad (1)$$

This property guarantees that the predicted distribution is statistically consistent with the observed outcomes. This condition holds if $\hat{F}_{Y|X}$ matches the true conditional CDF. In practice, the deviation of the PIT distribution from uniformity can be quantified using the *probabilistic calibration error* (PCE), defined as

$$\text{PCE}(F_\theta, \mathcal{D}) = \frac{1}{M} \sum_{j=1}^M \left| \alpha_j - \hat{F}_Z(\alpha_j) \right|, \quad (2)$$

where $\{\alpha_j\}_{j=1}^M$ is a grid of quantile levels such that $\alpha_j \in [0, 1]$, and $\hat{F}_Z$ is the empirical CDF of the PIT values $Z_i = \hat{F}_{Y|X=X_i}(Y_i)$, given by $\hat{F}_Z(\alpha) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(Z_i \leq \alpha)$. Although this nonparametric estimator is effective for evaluation purposes, its non-differentiability prevents its direct application during training.

Regularization for univariate calibration. To improve the calibration of probabilistic models, regularization-based approaches add a calibration-specific penalty term to the training objective, explicitly encouraging the PIT values to follow a uniform distribution (Wilks 2018; Dheur and Taieb 2023). These methods aim to improve calibration, potentially at the expense of sharpness, with the trade-off controlled by a regularization hyperparameter.

Among such approaches, Dheur and Taieb (2023) introduced a differentiable **PCE-KDE** regularizer, which smooths the empirical CDF of the PIT values using a logistic kernel density estimator (KDE). Given PIT values $Z_i = \hat{F}_{Y|X}(Y_i)$, the smoothed CDF at a grid point α_j is defined as:

$$\Phi_{\text{KDE}}(\alpha_j; \{Z_i\}_{i=1}^N) = \frac{1}{N} \sum_{i=1}^N \sigma(\tau(\alpha_j - Z_i)), \quad (3)$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, and $\tau > 0$ controls the smoothness of the approximation. The resulting regularization term is given by:

$$\mathcal{R}_{\text{PCE-KDE}} = \frac{1}{M} \sum_{j=1}^M \left| \alpha_j - \Phi_{\text{KDE}}(\alpha_j; \{Z_i\}_{i=1}^N) \right|^p, \quad (4)$$

where $p \geq 1$ determines the penalty's shape. Minimizing this term encourages the PIT distribution to align with the uniform distribution, thereby promoting probabilistic calibration during training.

Related Work

Univariate Calibration. A variety of methods have been proposed for univariate calibration, including post-hoc recalibration techniques (Kuleshov, Fenner, and Ermon 2018; Kuleshov and Deshpande 2021; Song et al. 2019) and regularization-based approaches (Zhao, Ma, and Ermon 2020; Feldman, Bates, and Romano 2021). Dheur and Taieb (2024) provide a unified perspective on these methods and introduce a training framework that integrates recalibration directly into the learning process. Tail-focused calibration has also gained attention, particularly through the use of weighted scoring rules and loss regularization (Wessel et al. 2025). More recently, hybrid strategies that combine refinement during training with post-hoc calibration have been shown to improve both sharpness and reliability (Berta et al. 2025).

Multivariate Calibration. Extending calibration to multivariate outputs is substantially more challenging due to the need to capture joint dependencies among target dimensions. One approach, known as *copula calibration*, evaluates the uniformity of the copula PIT, generalizing univariate rank histograms to the multivariate setting by assessing the joint CDF of the predicted distribution (Ziegel and Gneiting 2013). While conceptually appealing, no practical method currently exists to enforce copula calibration during training. An alternative line of work introduces *pre-rank functions*, which project multivariate forecast-observation pairs to scalar quantities before constructing rank histograms (Allen, Ziegel, and Ginsbourger 2023). These functions enable diagnostic assessment of specific forms of miscalibration but do not provide a mechanism for enforcing calibration under a given pre-rank. More recently, *HDR calibration* has been proposed to target high-density regions of the predictive distribution (Chung, Char, and Schneider 2024). It operates post-hoc by learning a mapping that resamples predictions to satisfy HDR calibration, followed by a correction step that updates the predictive model itself, though this step currently applies only to Gaussian outputs.

Multivariate calibration with pre-ranks

A closer examination of the multivariate calibration methods discussed in the previous section reveals that they can be interpreted within a unified framework based on *pre-rank functions*. These are univariate functionals $\rho : \mathcal{X} \times \mathbb{R}^D \rightarrow \mathbb{R}$ that map multivariate forecast-observation pairs to scalar values for calibration assessment. Each pre-rank highlights a specific structural aspect of the predictive distribution. Let $(X, Y) \sim F_{Y|X}$ and define

$$T = \rho(X, Y) \quad \text{and} \quad \hat{T} = \rho(X, \hat{Y}),$$

where $\hat{Y} \sim \hat{F}_{Y|X}$ denotes a sample from the predictive distribution.

Definition 2. $\hat{F}_{Y|X}$ is said to be **calibrated with respect to a pre-rank** ρ if $\hat{F}_{T|X}$ is PIT-calibrated (see Definition 1)

As shown in Chung, Char, and Schneider (2024), if the predictive distribution matches the true conditional distribution, i.e., $\hat{F}_{Y|X} = F_{Y|X}$, then calibration holds for any

choice of pre-rank ρ . In Table 1, we present several pre-rank functions previously introduced in the literature and considered in this work.

Pre-rank	Formula
Marginal	$\rho_{\text{marg}}^d(x, y) = y_d$
Location	$\rho_{\text{loc}}(x, y) = \frac{1}{D} \sum_{d=1}^D y_d$
Scale	$\rho_{\text{scale}}(x, y) = \frac{1}{D} \sum_{d=1}^D (y_d - \bar{y})^2$
Dependency	$\rho_{\text{dep}}(x, y; h) = -\frac{\gamma_y(h)}{s_y^2}$
HDR	$\rho_{\text{hdr}}(x, y) = \hat{f}_{Y X=x}(y)$
Copula	$\rho_{\text{cop}}(x, y) = \hat{F}_{Y X=x}(y)$

Table 1: Types of pre-rank functions considered in this work.

The marginal pre-rank assesses calibration along individual dimensions by extracting the d -th coordinate for each $d \in \{1, \dots, D\}$. The location pre-rank averages across dimensions to evaluate global bias, while the scale pre-rank measures the overall spread. The dependency pre-rank captures structural dependencies via a normalized variogram. For $h \in \{1, \dots, D-1\}$ it is defined as $\gamma_y(h) = \frac{1}{2(D-h)} \sum_{d=1}^{D-h} |y_d - y_{d+h}|^2$, where s_y^2 is a variance across dimensions and acts as a normalizer. The HDR pre-rank (Chung, Char, and Schneider 2024) adopts a likelihood-based perspective by evaluating the predicted density at the observed outcome. The Copula pre-rank, on the other hand, evaluates the predicted CDF at the observation, capturing the structure of the joint predictive distribution.

These techniques offer complementary insights into the quality of probabilistic predictions by evaluating how well the model captures structural or distributional aspects of the output. However, we emphasize that these pre-rank functions are primarily diagnostic in nature. The existing literature does not offer a principled way to incorporate them into the training process to enforce multivariate calibration. This is precisely the gap our work addresses.

An Experimental Study of Multivariate Calibration

We conduct a large-scale experimental study to assess the multivariate calibration of (unregularized) multi-output regression models using a set of pre-rank functions introduced in the previous section.

Benchmark Datasets. Our experiments are performed on 18 real-world multi-output regression datasets drawn from prior work (Feldman, Bates, and Romano 2022; Wang et al. 2022; Camehl, Fok, and Gruber 2025). These datasets are widely used in the literature on multivariate calibration (Chung, Char, and Schneider 2024), conformal prediction (Dheur et al. 2025; Guan 2021), and uncertainty quantification (Angelopoulos et al. 2020), and serve as a standard benchmark for evaluating calibration methods. We include only datasets with at least 400 training instances and follow the same preprocessing and train-validation-test splitting procedure as in Dheur et al. (2025). The selected datasets vary in size, containing between 424 and 406,440 training examples. The number of input features L ranges from 1 to

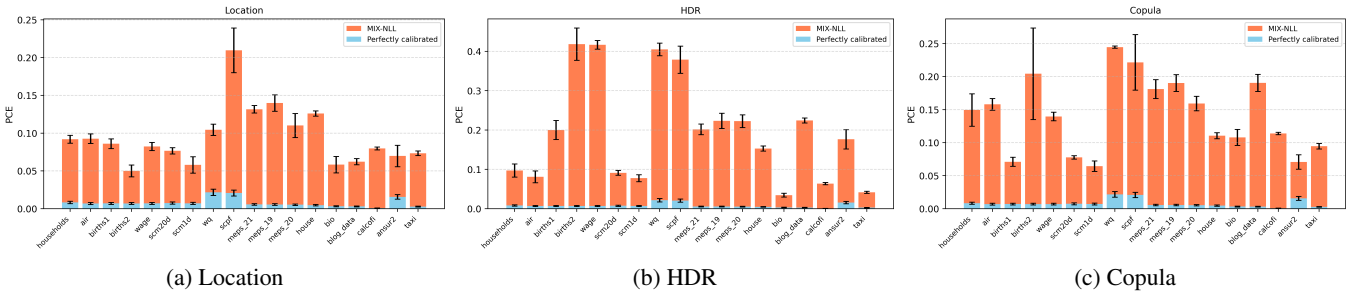


Figure 1: PCE values with respect to (a) Location (b) HDR and (c) Copula pre-ranks averaged over five runs across 18 benchmark datasets using the MIX-NLL baseline. Blue bars indicate reference PCE values from a simulated perfectly calibrated model.

279, and the number of output variables D ranges from 2 to 16.

Neural probabilistic regression model. Our base probabilistic predictor models a conditional predictive distribution as a mixture of K multivariate Gaussian components, where all parameters are generated by a hypernetwork. For each input $x \in \mathcal{X}$ and each mixture component $k \in [K]$, the network predicts the mixture weight $\pi_k(x)$, the mean vector $\mu_k(x) \in \mathbb{R}^D$, and the lower triangular Cholesky factor $L_k(x)$. The covariance matrix is then computed as $\Sigma_k(x) = L_k(x)L_k(x)^\top$, ensuring positive semi-definiteness by construction. The resulting conditional density takes the form: $\hat{f}_{Y|X=x} = \sum_{k=1}^K \pi_k(x) \mathcal{N}(\cdot | \mu_k(x), \Sigma_k(x))$ where $\pi_k(x) \geq 0$ and $\sum_{k=1}^K \pi_k(x) = 1$. We train this model using the NLL scoring rule and refer to this baseline as MIX-NLL. Further architectural and training details are provided in the Experiments section.

Results. Figure 1 reports the test PCE values for the location, HDR, and copula pre-rank functions, averaged over five independent runs (corresponding to different train-validation-test splits) on each of the 18 benchmark datasets using the MIX-NLL model. For reference, we also simulate ideal PCE scores by sampling from a perfectly calibrated model; these reference values are shown in blue. Due to space constraints, we display results for only a subset of pre-rank functions; figures for the remaining ones are provided in the Appendix. As shown, the MIX-NLL model exhibits substantial miscalibration across all pre-ranks and the majority of datasets.

We assess the significance of PCE values by generating 5×10^4 samples of uniformly distributed PITs to approximate the null distribution under perfect calibration for each dataset and pre-rank. One-sided p-values (Holm-corrected) show that all deviations are statistically significant, confirming systematic miscalibration (see Appendix).

In summary, these results highlight that despite being trained with a strictly proper scoring rule, MIX-NLL exhibits significant miscalibration across multiple pre-rank functions on standard benchmarks. In the following section, we investigate how calibration can be improved for these pre-ranks.

A Pre-rank Regularization Framework

Although proper scoring rules are designed to reward calibration, minimizing them during training does not guarantee that the resulting models will be calibrated. Under model misspecification, even strictly proper scoring rules may favor sharp yet miscalibrated predictions, as they do not explicitly penalize miscalibration (Bröcker 2008).

In this section, we introduce a training strategy that explicitly enforces calibration by augmenting the loss function with a calibration-specific regularization term. Building on the pre-rank functions introduced earlier, we leverage *projected PITs* to reduce the multivariate calibration problem to a collection of univariate calibration tasks.

Calibration of Projected PITs. As stated in Definition 2, assessing calibration requires access to the conditional CDF $\hat{F}_{T|X}$. Since this CDF is typically unavailable in closed form, we approximate it empirically. For a given test point (X_i, Y_i) , we draw S samples $\hat{Y}_1, \dots, \hat{Y}_S \sim \hat{F}_{Y|X=X_i}$, and compute the projected values:

$$T_i = \rho(X_i, Y_i) \quad \text{and} \quad \hat{T}_s = \rho(X_i, \hat{Y}_s) \quad \text{for } s = 1, \dots, S.$$

The conditional CDF $\hat{F}_{T|X=X_i}$ is then estimated using a smoothed indicator function:

$$\hat{F}_{T|X=X_i}(t) = \frac{1}{S} \sum_{s=1}^S \mathbf{1}_{\tau}(\hat{T}_s \leq t), \quad (5)$$

where $\mathbf{1}_{\tau}(x \leq y) = \sigma(\tau(y - x))$, and $\sigma(z) = \frac{1}{1 + e^{-z}}$ is the sigmoid function with temperature parameter τ . The *projected PIT* is defined as the value of this estimated CDF evaluated at the true projected target:

$$Z = \hat{F}_{T|X}(T). \quad (6)$$

Under perfect calibration for the pre-rank function ρ , the projected PIT values Z should follow a uniform distribution in $[0, 1]$.

As in equation 2, under a given pre-rank function ρ , the PCE can be used to quantify the deviation of projected PIT values from uniformity. This extends the univariate PCE formulation to the multivariate setting by applying it to any scalar projection of the multivariate predictions, thereby enabling the assessment of calibration with respect to a chosen

pre-rank. Note, however, that the empirical CDF of the projected PITs is not differentiable and therefore cannot be used directly in gradient-based training. To address this, we rely on differentiable approximations that enable calibration to be enforced during model training.

Pre-Rank calibration via Regularization. Recall from equation 6 that the projected PIT variable Z depends on a chosen pre-rank function ρ . To encourage calibration with respect to ρ , we define a differentiable regularizer based on the PCE-KDE expression in equation 4, using the projected PIT values.

Following Wilks (2018) and Dheur and Taieb (2024), we augment the training loss with a differentiable penalty that steers the model toward improved calibration during training. Specifically, the augmented objective is:

$$\mathcal{L}(\theta; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^N S(F_\theta(x_i), y_i) + \lambda \mathcal{R}_{\text{PCE-KDE}}(\theta; \mathcal{D}; \rho), \quad (7)$$

where S is a strictly proper scoring rule and $\lambda \geq 0$ controls the strength of the calibration regularization. The case $\lambda = 0$ recovers standard unregularized training, while increasing λ prioritizes calibration, potentially at the expense of predictive accuracy. Note that the regularizer $\mathcal{R}_{\text{PCE-KDE}}(\theta; \mathcal{D}; \rho)$ is specific to the chosen pre-rank function ρ . We refer to the resulting method as **pre-rank**.

Marginal and Pre-rank Calibration. Calibration with respect to an arbitrary pre-rank does not necessarily imply marginal calibration for each output dimension. However, ensuring marginal calibration is crucial, as any multivariate distribution can be decomposed into its marginal distributions and a dependence structure, according to Sklar’s theorem (Sklar 1959). An important exception is the copula pre-rank, which is designed to capture both marginal and joint miscalibration. Since this property does not hold for many commonly used pre-ranks, we propose to explicitly account for marginal calibration alongside any chosen pre-rank. To this end, we define a combined regularizer:

$$\frac{1}{D} \sum_{d=1}^D \mathcal{R}_{\text{PCE-KDE}}(\theta; \mathcal{D}; \rho_{\text{marg}}^d) + \mathcal{R}_{\text{PCE-KDE}}(\theta; \mathcal{D}; \rho), \quad (8)$$

where ρ_{marg}^d denotes the marginal pre-rank function for the d -th output dimension.

As before, any training loss can be augmented with this combined regularizer. We refer to the resulting model variant as **marginal+pre-rank**. This formulation is designed to enforce marginal calibration without compromising, and potentially enhancing, calibration along the selected pre-rank direction.

A PCA-Based Pre-rank and Regularizer. Given the multivariate nature of the output, we propose a novel pre-rank function based on principal component analysis (PCA). The goal is to assess calibration along the directions of highest variance in the predictive distribution. Specifically, the PCA pre-rank projects the output onto the top principal components of the model’s predictive covariance, yielding the

following function:

$$\rho_{\text{pca}}^d(x, y) = y \cdot V_d(x),$$

where $V_d(x) \in \mathbb{R}^D$, for $d \in \{1, \dots, D\}$, denotes the d -th principal component of the covariance matrix associated with the predicted conditional distribution $\hat{F}_{Y|X=x}$. These components are obtained by sampling from the model’s predictive distribution and performing PCA on the resulting samples.

While our PCA pre-rank can be treated like any other pre-rank function, it offers the additional advantage of dimensionality reduction. Its computational complexity is $O(SD^2 + D^3)$, but when only the top principal components d^* are retained, the combined PCA + pre-rank cost scales with d^* instead of D . This is particularly beneficial in high-dimensional settings, where evaluating calibration across all marginals can be computationally intensive and statistically unstable.

To this end, we project the outputs onto the top d^* principal components that explain a large proportion of the predictive variance (e.g., 80%). We then compute PCEs along these components and combine them with the regularization term from an arbitrary pre-rank ρ , yielding the following combined regularizer:

$$\frac{1}{d^*} \sum_{d=1}^{d^*} \mathcal{R}_{\text{PCE-KDE}}(\theta; \mathcal{D}; \rho_{\text{pca}}^d) + \mathcal{R}_{\text{PCE-KDE}}(\theta; \mathcal{D}; \rho).$$

We refer to this approach as **PCA+pre-rank**.

To clarify when different pre-rank calibration notions are interchangeable, we provide a sufficient condition under which calibration with respect to one pre-rank function implies calibration with respect to another.

Equivalence of Pre-Rank Calibration. Let ρ_1 and ρ_2 be two projection functions. We say that the calibration criteria associated with ρ_1 and ρ_2 are *equivalent* if a model is calibrated with respect to ρ_1 if and only if it is calibrated with respect to ρ_2 . The following proposition characterizes a sufficient condition for such equivalence:

Proposition 1. For every fixed $x \in \mathbb{R}^L$, the function $y \mapsto \rho_2(x, y)$ must be a strictly monotonic bijective transformation of $y \mapsto \rho_1(x, y)$. That is, there exists a strictly increasing or decreasing bijection h_x such that for all $y \in \mathbb{R}^D$,

$$\rho_2(x, y) = h_x(\rho_1(x, y)).$$

The full proof is provided in the Appendix. This result shows that strictly monotonic transformations of projection functions preserve the distribution of PIT values, and therefore, the notion of calibration. However, such conditions are rarely met in practice, and different pre-rank functions often lead to distinct, potentially incompatible calibration assessments.

Experiments

We extend our earlier empirical analysis to evaluate the effectiveness of the proposed pre-rank regularization framework. Specifically, we train the MIX-NLL model with a

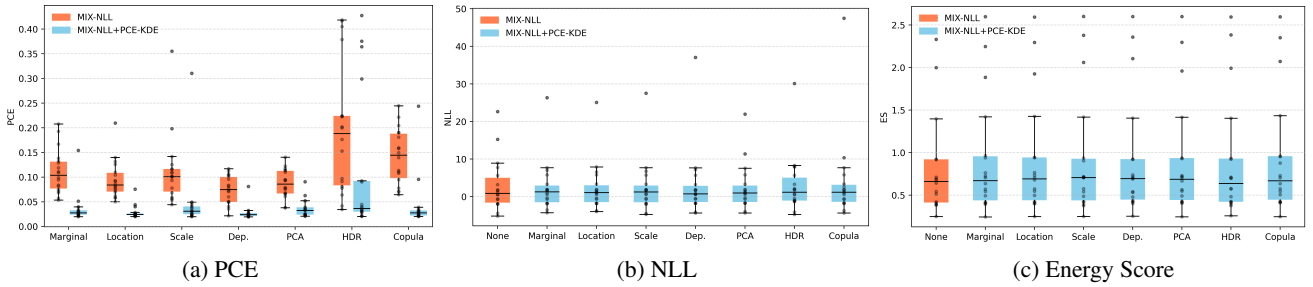


Figure 2: Performance on 18 real multivariate benchmark datasets. Orange: MIX-NLL (no regularization). Blue: MIX-NLL+PCE-KDE (proposed). Metrics are calculated across seven pre-rank functions, and averaged over five runs. In subplots (b) and (c), the “None” box refers to the unregularized MIX-NLL trained without pre-rank.

PCE-KDE regularizer applied to the following pre-rank functions: (1) marginal, (2) location, (3) scale, (4) dependency, (5) PCA, (6) HDR, and (7) copula. We refer to this model as **MIX-NLL+PCE-KDE** trained on **pre-rank**. We then compute the PCE values on the test set and compare them to those obtained from the unregularized MIX-NLL baseline. All experiments are conducted using the same 18 benchmark datasets introduced earlier.

Metrics. We evaluate model performance using three metrics: PCE (as a measure of calibration), negative log-likelihood (NLL), and energy score (ES). Details on the empirical computation of ES are provided in the Appendix.

Hyperparameters. For the MIX-NLL baseline, we use a mixture of $K = 5$ multivariate Gaussian components. The neural network consists of three fully connected layers with 100 hidden units each, ReLU activations, and is trained using the Adam optimizer with a learning rate of 10^{-4} . To compute the PCE-KDE regularizer, we estimate the projected PITs $\hat{F}_{T|X}(T)$ using $S = 100$ samples drawn from the predictive distribution with the parameters set to $p = 1$ and $M = 100$. The temperature parameter τ in the smoothed indicator function is set to 100, following prior work in Dheur and Taieb (2023). The regularization strength λ in equation 7 controls the degree of calibration enforcement with respect to the chosen pre-rank. As observed in prior work (Karandikar et al. 2021; Wessel et al. 2025), increasing λ typically improves calibration (lower PCE) but may degrade predictive performance (higher NLL or ES). Following the tuning strategy used in Karandikar et al. (2021) and Dheur and Taieb (2023), we select λ to minimize PCE while ensuring that ES does not increase by more than 10% relative to the best ES obtained when $\lambda = 0$. This strategy allows us to improve calibration without sacrificing predictive accuracy. The optimal λ is tuned on validation set and selected from $\{0, 0.01, 0.1, 1, 5, 10\}$ for each (dataset, pre-rank) pair. The exact values of selected λ are reported in the Appendix.

Results

Figure 2a compares the test PCE of the unregularized MIX-NLL model with the regularized version, MIX-NLL+PCE-KDE, where calibration is explicitly enforced with respect

to each pre-rank function. As expected, regularization substantially reduces the PCE for the corresponding pre-rank. For all pre-ranks, the median PCE across datasets is consistently lower after regularization. Additionally, the distribution of PCE values (illustrated via box plots) becomes noticeably tighter, indicating that the regularization leads to more consistent calibration improvements across datasets. Full results, averaged over five runs for each dataset and pre-rank, are provided in the Appendix.

Despite the overall improvements, calibration remains challenging for a few datasets with initially high PCE values, particularly under the HDR pre-rank. As shown in Figure 2a, four datasets exhibit little to no improvement in PCE when regularized with HDR. This aligns with the limitation highlighted by Chung, Char, and Schneider (2024): when the model’s predictive distribution poorly approximates the true one, HDR recalibration struggles to recover the underlying dependency structure among target variables. Consequently, the effectiveness of HDR as a pre-rank is highly sensitive to model specification. In these cases, the underlying Mixture of Gaussians model may be misaligned with the data distribution.

Figures 2b and 2c show the NLL and ES values after applying pre-rank-based regularization, alongside their values without regularization (denoted “None”). As the results indicate, regularization does not significantly degrade predictive performance. This demonstrates that enforcing calibration through pre-rank regularization maintains predictive quality. Importantly, the regularization strength λ is selected to control increases in ES, ensuring that improvements in calibration do not come at the expense of accuracy.

Marginal and Pre-rank Calibration. The reliability plots in Figure 3 show that regularizing solely with respect to a specific pre-rank improves calibration for that pre-rank, but not necessarily for the marginal distributions. Among the three examples shown, only the copula pre-rank also improves marginal calibration, as it is designed to capture both marginal and joint structure. A similar effect is observed with the PCA pre-rank (see Appendix), which improves marginal calibration by projecting prediction-observation pairs onto all principal components and averaging the resulting PCEs, acting as a marginal pre-rank in a rotated space. In contrast, HDR-only regularization shows minimal improve-

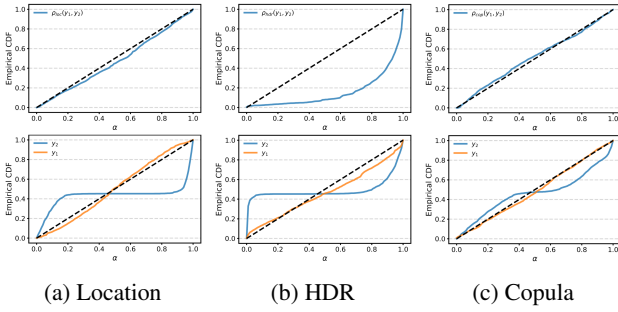


Figure 3: Reliability plots on *wage* dataset using MIX-NLL + PCE-KDE on prerank. Top row: calibration curves with respect to (a) Location, (b) HDR, and (c) Copula preranks. Bottom row: corresponding marginal calibration curves.

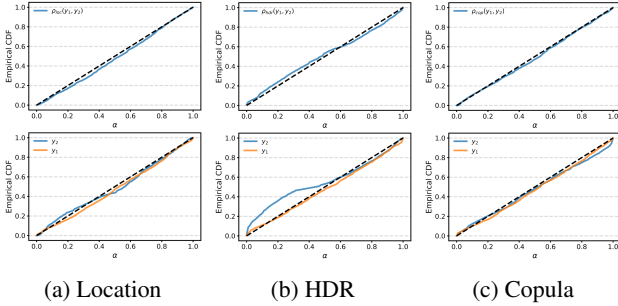


Figure 4: Reliability plots on *wage* dataset using MIX-NLL + PCE-KDE on marginal+pre-rank. Top row: calibration curves with respect to (a) Location, (b) HDR, and (c) Copula pre-ranks. Bottom row: corresponding marginal calibration curves.

ment on the *wage* dataset (Figure 3b), likely due to its dependence on the quality of the model’s predictive distribution. By comparison, Figure 4b demonstrates that combining marginal and pre-rank regularization leads to improvements in both HDR and marginal calibration. This is likely because the marginal PCE provides a complementary signal that is less sensitive to model misspecification, thereby supporting more reliable calibration across both marginal and structured components.

PCA and Pre-rank Calibration. Table 2 reports both marginal and pre-rank PCEs for four training variants: None (no regularization), pre-rank (regularized on a certain pre-rank only), marginal+pre-rank, and PCA+pre-rank. Results are shown for the *scml*d dataset with $D = 16$ target dimensions. As expected, the marginal+pre-rank variant achieves the lowest marginal PCEs. However, PCA+pre-rank performs comparably well: even after reducing the dimensionality from 16 to just 3 principal components, its marginal PCEs remain substantially lower than those of the unregularized (None) model. In terms of pre-rank PCEs, marginal+pre-rank again achieves the best results, but the differences are modest. On average, PCA+pre-rank increases the pre-rank PCE by no more than 5% compared to marginal+pre-rank, while still improving significantly over

the None baseline. This indicates that PCA+pre-rank offers a scalable and effective alternative, achieving near-comparable calibration performance using only a few informative projections. Additional results for datasets with $D \geq 4$ are provided in the Appendix.

Method	Marg	Loc	Scale	Dep	PCA	HDR	Cop
None	0.054	0.058	0.108	0.084	0.038	0.078	0.064
Marg	0.025	0.028	0.074	0.063	0.038	0.08	0.040
Loc	0.030	0.024	0.087	0.072	0.040	0.080	0.046
Scale	0.047	0.059	0.037	0.038	0.038	0.075	0.055
Dep	0.059	0.071	0.078	0.020	0.043	0.081	0.064
PCA	0.029	0.023	0.079	0.049	0.035	0.085	0.04
HDR	0.059	0.069	0.113	0.089	0.042	0.091	0.069
Cop	0.035	0.04	0.085	0.066	0.04	0.083	0.035
Marg+loc	0.024	0.022	0.074	0.05	0.036	0.073	0.038
Marg+scale	0.026	0.031	0.054	0.05	0.038	0.082	0.044
Marg+dep	0.025	0.030	0.071	0.021	0.037	0.085	0.041
Marg+HDR	0.033	0.037	0.094	0.059	0.035	0.087	0.042
Marg+Cop	0.024	0.027	0.076	0.057	0.036	0.078	0.028
PCA+loc	0.030	0.023	0.081	0.074	0.040	0.093	0.043
PCA+scale	0.032	0.032	0.052	0.055	0.036	0.078	0.041
PCA+dep	0.030	0.025	0.074	0.022	0.036	0.081	0.039
PCA+HDR	0.043	0.044	0.115	0.087	0.038	0.095	0.06
PCA+Cop	0.036	0.043	0.077	0.105	0.046	0.101	0.042

Table 2: PCE values averaged over five runs from four model variants: **None** (no regularization), **pre-rank** (regularized on certain pre-rank), **marg+pre-rank**, and **PCA+pre-rank**. PCE values are shown across different pre-ranks.

Conclusion

Multivariate calibration is often assessed using pre-rank functions—projections that reduce prediction-observation pairs to univariate summaries, such as marginal, location, scale, or dependency-based mappings. In a large-scale empirical study on 18 real-world regression datasets, we show that a standard probabilistic predictor, despite being trained with a strictly proper scoring rule, is consistently miscalibrated across all pre-ranks.

To address this, we propose a differentiable regularization framework that enforces calibration during training by penalizing the deviation between quantile levels and the empirical CDF of projected PITs. The method integrates seamlessly with any scoring-rule-based objective and can be extended to jointly enforce marginal and pre-rank calibration.

We also introduce a PCA-based pre-rank that projects predictions onto principal directions of variance, enabling effective calibration in a lower-dimensional space. Despite using only a few components, PCA+pre-rank achieves calibration performance close to marginal+pre-rank.

Empirical results show that our approach consistently improves calibration without compromising predictive accuracy. Overall, this work offers a practical strategy for enforcing multivariate calibration and opens avenues for integrating projection-based regularization into model training.

References

- Allen, S.; Ziegel, J.; and Ginsbourger, D. 2023. Assessing the calibration of multivariate probabilistic forecasts. *arXiv [stat.ME]*.
- Angelopoulos, A. N.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty Sets for Image Classifiers using Conformal Prediction. *ArXiv*, abs/2009.14193.
- Berta, E.; Holzmüller, D.; Jordan, M. I.; and Bach, F. 2025. Rethinking Early Stopping: Refine, Then Calibrate. *arXiv preprint*.
- Bröcker, J. 2008. Reliability, sufficiency, and the decomposition of proper scores. *arXiv [physics.ao-ph]*.
- Camehl, A.; Fok, D.; and Gruber, K. 2025. On superlevel sets of conditional densities and multivariate quantile regression. *Journal of Econometrics*, 249: 105807.
- Chung, Y.; Char, I.; and Schneider, J. 2024. Sampling-based Multi-dimensional Recalibration. In *Forty-first International Conference on Machine Learning*.
- de Lima Silva, P. C.; Sadaei, H. J.; Ballini, R.; and Guimarães, F. G. 2020. Probabilistic Forecasting With Fuzzy Time Series. *IEEE Transactions on Fuzzy Systems*, 28(8): 1771–1784.
- Dheur, V.; Fontana, M.; Estievenart, Y.; Desobry, N.; and Taieb, S. B. 2025. A unified comparative study with generalized conformity scores for multi-output conformal regression. *arXiv [stat.ML]*.
- Dheur, V.; and Taieb, S. B. 2023. A large-scale study of probabilistic calibration in neural network regression. *arXiv [cs.LG]*.
- Dheur, V.; and Taieb, S. B. 2024. Probabilistic calibration by design for neural network regression. *arXiv [cs.LG]*.
- Feldman, S.; Bates, S.; and Romano, Y. 2021. Improving conditional coverage via orthogonal quantile regression. *arXiv [cs.LG]*.
- Feldman, S.; Bates, S.; and Romano, Y. 2022. Calibrated Multiple-Output Quantile Regression with Representation Learning. *arXiv*:2110.00816.
- Gneiting, T.; and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 102(477): 359–378.
- Guan, L. 2021. Localized Conformal Prediction: A Generalized Inference Framework for Conformal Prediction. *Biometrika*.
- Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; and Gaidon, A. 2019. 3D packing for self-supervised monocular depth estimation. *arXiv [cs.CV]*.
- Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; Kim, R.; Raman, R.; Nelson, P. C.; Mega, J. L.; and Webster, D. R. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410.
- Karandikar, A.; Cain, N.; Tran, D.; Lakshminarayanan, B.; Shlens, J.; Mozer, M. C.; and Roelofs, B. 2021. Soft Calibration Objectives for Neural Networks. *arXiv*:2108.00106.
- Krzysztofowicz, R.; and Evans, W. B. 2008. Probabilistic Forecasts from the National Digital Forecast Database. *Weather and Forecasting*, 23(2): 270 – 289.
- Kuleshov, V.; and Deshpande, S. 2021. Calibrated and sharp uncertainties in deep learning via density estimation. *ICML*, 162: 11683–11693.
- Kuleshov, V.; Fenner, N.; and Ermon, S. 2018. Accurate uncertainties for deep learning using calibrated regression. *ICML*, abs/1807.00263: 2796–2804.
- Murphy, A. H.; and Winkler, R. L. 1984. Probability Forecasting in Meteorology. *Journal of the American Statistical Association*, 79(387): 489–500.
- Sklar, M. 1959. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8: 229–231.
- Song, H.; Diethe, T.; Kull, M.; and Flach, P. 2019. Distribution Calibration for Regression. *arXiv [stat.ML]*.
- Wang, Z.; Gao, R.; Yin, M.; Zhou, M.; and Blei, D. M. 2022. Probabilistic Conformal Prediction Using Conditional Random Samples. *arXiv*:2206.06584.
- Wessel, J. B.; Schillinger, M.; Kwasniok, F.; and Allen, S. 2025. Enforcing tail calibration when training probabilistic forecast models. *arXiv [stat.AP]*.
- Wilks, D. S. 2018. Enforcing calibration in ensemble post-processing: Enforcing Calibration in Ensemble Postprocessing. *Q. J. R. Meteorol. Soc.*, 144(710): 76–84.
- Zhao, S.; Ma, T.; and Ermon, S. 2020. Individual calibration with randomized forecasting. *arXiv [stat.ML]*.
- Ziegel, J. F.; and Gneiting, T. 2013. Copula Calibration. *arXiv [stat.ME]*.
- Önkal, D.; and Muradoğlu, G. 1994. Evaluating probabilistic forecasts of stock prices in a developing stock market. *European Journal of Operational Research*, 74(2): 350–358. Financial Modelling.

Appendix

A. Proofs

Equivalence of Pre-Rank Calibration

Proposition 2. For every fixed $x \in \mathbb{R}^L$, the function $y \mapsto \rho_2(x, y)$ must be a strictly monotonic bijective transformation of $y \mapsto \rho_1(x, y)$. That is, there exists a strictly increasing or decreasing bijection h_x such that for all $y \in \mathbb{R}^D$,

$$\rho_2(x, y) = h_x(\rho_1(x, y)).$$

Proof. Fix $x \in \mathbb{R}^L$ and define $T_1 = \rho_1(x, Y)$ and $T_2 = \rho_2(x, Y) = h_x(T_1)$, where h_x is a strictly monotonic bijection. Let $\hat{F}_{T_1|X=x}$ and $\hat{F}_{T_2|X=x}$ denote the empirical conditional CDFs of T_1 and T_2 , respectively, estimated using the same sample of predicted values $\{\hat{Y}_i\}_{i=1}^{N'}$ drawn from the predictive distribution $\hat{F}_{Y|X=x}$.

As explained in the background section, we estimate these conditional CDFs using the empirical estimator. Since this construction is used solely for evaluation, differentiability of the CDF is not required. Then for any $t \in \mathbb{R}$,

$$\begin{aligned} \hat{F}_{T_2|X=x}(t) &= \frac{1}{N'} \sum_{i=1}^{N'} \mathbf{1}_{\tau}(\rho_2(x, \hat{Y}_i) \leq t) \\ &= \frac{1}{N'} \sum_{i=1}^{N'} \mathbf{1}_{\tau}(\rho_1(x, \hat{Y}_i) \leq h_x^{-1}(t)) \\ &= \hat{F}_{T_1|X=x}(h_x^{-1}(t)). \end{aligned}$$

Since $T_2 = h_x(T_1)$ and

$$\hat{F}_{T_2|X=x}(t) = \hat{F}_{T_1|X=x}(h_x^{-1}(t)),$$

we have:

$$\hat{F}_{T_2|X=x}(T_2) = \hat{F}_{T_1|X=x}(h_x^{-1}(T_2)) = \hat{F}_{T_1|X=x}(T_1),$$

where we used the fact that $T_1 = h_x^{-1}(T_2)$ by construction. It follows that the PIT value computed under ρ_2 coincides with the one computed under ρ_1 :

$$U_2 := \hat{F}_{Z_2|X=x}(Z_2) = \hat{F}_{Z_1|X=x}(Z_1) =: U_1.$$

It follows that U_1 and U_2 have the same distribution. In particular,

$$U_2 \sim \mathcal{U}[0, 1] \iff U_1 \sim \mathcal{U}[0, 1],$$

which establishes the equivalence of the two calibration criteria under the assumed transformation.

B. Additional computational details

Practical note on Copula-based Pre-Rank. When using copula-based pre-ranks, one requires access to the joint CDF $\hat{F}_{Y|X}(y)$, i.e., the probability that all components of Y are less than or equal to y given X . However, many models provide only the conditional density $\hat{f}_{Y|X}$.

We approximate the joint CDF via Monte Carlo sampling. Given input X_i and target Y_i , we draw S samples $\hat{Y}_{i,1}, \dots, \hat{Y}_{i,S} \sim \hat{f}_{Y|X=X_i}$ and estimate:

$$\hat{F}_{Y|X=X_i}(Y_i) \approx \frac{1}{S} \sum_{s=1}^S \mathbf{1}\{\hat{Y}_{i,s} \leq Y_i\}, \quad (9)$$

where the indicator $\mathbf{1}\{\hat{Y}_{i,s} \leq y_i\}$ is true if and only if $\hat{Y}_{i,s}^{(d)} \leq y_i^{(d)}$ for all components $d = 1, \dots, D$.

Since the indicator function is not differentiable, we replace it with a smooth approximate using the sigmoid function $\sigma(z) = 1/(1+e^{-z})$ and a temperature parameter $\tau > 0$. This gives:

$$\hat{F}_{Y|X=X_i}(Y_i) \approx \frac{1}{S} \sum_{s=1}^S \prod_{d=1}^D \sigma\left(\tau(Y_i^{(d)} - \hat{Y}_{i,s}^{(d)})\right), \quad (10)$$

where $y_i^{(d)}$ and $\hat{Y}_{i,s}^{(d)}$ denote the d -th components of the vectors y_i and $\hat{Y}_{i,s}$, respectively. The product over dimensions enforces that all components of $\hat{Y}_{i,s}$ fall below the threshold y_i , mimicking the joint indicator condition.

This smooth approximation is fully differentiable with respect to the model parameters (via the samples $\hat{Y}_{i,s}$), and thus compatible with gradient-based optimization routines such as backpropagation.

Empirical Calculation of Energy Score We use Energy Score (ES) as a scoring rule metric to evaluate our model performance. ES generalizes Continuous Ranked Probability Score (CRPS) to multivariate settings and is computed empirically as:

$$\text{ES}(\hat{F}, y) = \frac{1}{G} \sum_{i=1}^G \|\hat{Y}_i - y\| - \frac{1}{2G^2} \sum_{i=1}^G \sum_{j=1}^G \|\hat{Y}_i - \hat{Y}_j\| \quad (11)$$

where $\{\hat{Y}_i\}_{i=1}^G \sim \hat{F}_{Y|X}$ are G samples drawn from the predictive distribution. We set $G = 100$ in all experiments.

C. Detailed Hypothesis Test Results

Figure 5 shows PCE values for the marginal, scale, dependency, and PCA pre-ranks (excluded from the main paper), averaged over five independent runs on 18 real benchmark datasets using the MIX-NLL model. Simulated PCE scores from a perfectly calibrated model are shown in blue.

Distribution of the Test Statistic. To assess the statistical significance of observed PCE values, we estimate the null distribution of the average PCE under perfect calibration for each dataset and pre-rank. For every dataset, 5×10^4 samples of the test statistic are generated by simulating independent PIT values uniformly in $[0, 1]$, matching the test set size. This captures the variability of the mean PCE expected under ideal calibration.

One-sided p -values are computed as the proportion of simulated PCE values exceeding the observed PCE, and Holm correction is applied to control the family-wise error rate across datasets and pre-ranks. After correction, perfect calibration is rejected for all combinations, indicating systematic miscalibration.

Figures 6 - 12 show histograms of the null distributions of the average (over 5 runs) PCE for each dataset and pre-rank with the corresponding observed averages. In many cases, the observed average PCE lies deep in the right tail of the null distribution; for several datasets, it even exceeds all 10^4 simulated averages, demonstrating strong deviations from perfect calibration.

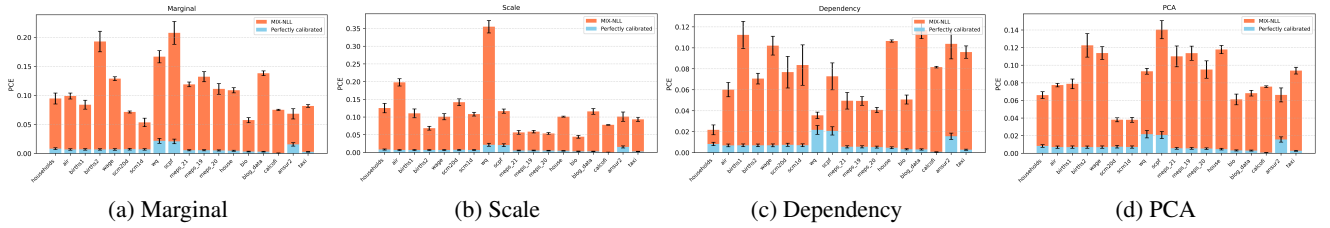


Figure 5: PCE values with respect to (a) Marginal (b) Scale (c) Dependency and (d) PCA pre-ranks averaged over five runs across 18 benchmark datasets using the MIX-NLL baseline. Blue bars indicate reference PCE values from a simulated perfectly calibrated model.

D. Hyperparameters

As described in the main paper, we select the λ that minimizes PCE while ensuring that ES does not increase by more than 10% relative to the reference ES from the best epoch of the model trained with $\lambda = 0$. This tuning is performed separately for each dataset and pre-rank pair. We select λ on validation set from $\{0, 0.01, 0.1, 1, 5, 10\}$. Table 3 shows the selected λ for each dataset-pre-rank pair. Notably, the majority of selected values are large, often $\lambda = 10$, suggesting that future work could explore larger values or employ more sophisticated tuning strategies such as Bayesian Optimization.

Datasets	Marginal	Loc.	Scale	Dep.	PCA	HDR	Copula
households	10.0	10.0	10.0	5.0	10.0	10.0	5.0
air	10.0	10.0	10.0	10.0	10.0	10.0	5.0
births1	10.0	10.0	10.0	10.0	10.0	5.0	10.0
births2	10.0	10.0	5.0	5.0	10.0	0.01	10.0
wage	10.0	10.0	5.0	10.0	5.0	1.0	10.0
scm20d	10.0	10.0	10.0	10.0	5.0	10.0	0.01
scm1d	10.0	10.0	10.0	10.0	10.0	0.1	10.0
wq	5.0	10.0	10.0	5.0	5.0	5.0	10.0
scpf	5.0	10.0	10.0	0.0	5.0	1.0	10.0
meps21	5.0	5.0	5.0	10.0	10.0	10.0	5.0
meps19	5.0	10.0	1.0	10.0	10.0	10.0	10.0
meps20	5.0	10.0	1.0	1.0	10.0	10.0	10.0
house	5.0	10.0	5.0	10.0	5.0	5.0	10.0
bio	5.0	10.0	10.0	10.0	5.0	10.0	5.0
blog data	10.0	10.0	10.0	10.0	10.0	10.0	10.0
calcofi	10.0	5.0	10.0	10.0	10.0	10.0	5.0
ansur2	10.0	5.0	5.0	10.0	10.0	5.0	10.0
taxi	10.0	10.0	10.0	5.0	10.0	5.0	10.0

Table 3: Values of λ after hyperparameter tuning with each regularization and each pre-rank. The baseline used is MIX-NLL.

E. Detailed Results

Pre-rank Calibration. Table 10 reports the exact PCE values averaged over five runs for each pre-rank on which the MIX-NLL+PCE-KDE model was trained using the optimal λ . For comparison, we also include the PCE values computed with respect to each pre-rank for the baseline MIX-NLL model trained without regularization (see Table 9). Note that although the baseline model was not trained with respect to any specific pre-rank, we still evaluate its perfor-

mance on each pre-rank to highlight the benefit of regularization.

Marginal and Pre-rank Calibration. Figures 13–17 show reliability plots for the majority of real benchmark datasets using the MIX-NLL+PCE-KDE model trained with each of the pre-ranks. We display only a representative subset of datasets due to the similarity of plots. In each figure, the top row shows calibration curves with respect to the pre-rank used during training, while the bottom row shows marginal calibration plots from the same models.

We observe that in many cases, the top-row plots exhibit strong alignment between the empirical CDF and the quantile levels α , indicating effective calibration with respect to the training pre-rank. However, the bottom-row plots reveal that strong calibration with respect to the pre-rank does not necessarily lead to better marginal calibration.

In contrast, Figures 18–22 present the reliability diagrams for the same datasets using the MIX-NLL+PCE-KDE model trained with the marginal+pre-rank objective. These plots demonstrate that jointly enforcing calibration with respect to both the marginal and the pre-ranks consistently results in strong calibration across both aspects.

PCA and Pre-rank Calibration. In the main paper, we reported calibration results on the `scm1d` dataset with 16 target dimensions, comparing four training variants: None (no regularization), pre-rank regularization, marginal+pre-rank, and PCA+pre-rank. We extend these observations by presenting additional PCE scores for datasets with target dimension $D \geq 4$.

These results confirm that the PCA+pre-rank approach consistently maintains competitive calibration performance using only a small number of informative projections, offering a practical solution for higher-dimensional multivariate regression tasks. In general, PCA offers a good trade-off between calibration quality and computational cost: it often achieves marginal and pre-rank-specific calibration performance close to that of the marginal+pre-rank method, while being significantly more efficient to compute. The full set of results is reported in Tables 4 - 8.

Method	Marg	Loc	Scale	Dep	PCA	HDR	Cop
None	0.072	0.077	0.142	0.077	0.038	0.091	0.078
Marg	0.033	0.035	0.094	0.042	0.033	0.085	0.037
Loc	0.036	0.025	0.109	0.053	0.034	0.086	0.046
Scale	0.053	0.071	0.042	0.054	0.039	0.107	0.055
Dep	0.072	0.081	0.117	0.025	0.038	0.08	0.072
PCA	0.048	0.045	0.114	0.058	0.033	0.086	0.054
HDR	0.082	0.086	0.167	0.096	0.039	0.061	0.086
Cop	0.042	0.051	0.125	0.048	0.033	0.084	0.031
Marg+loc	0.029	0.024	0.101	0.045	0.032	0.084	0.038
Marg+scale	0.027	0.033	0.041	0.045	0.032	0.077	0.035
Marg+dep	0.037	0.041	0.085	0.026	0.033	0.081	0.038
Marg+HDR	0.041	0.044	0.123	0.044	0.032	0.088	0.042
Marg+Cop	0.031	0.035	0.096	0.038	0.032	0.086	0.028
PCA+loc	0.034	0.023	0.102	0.035	0.031	0.087	0.045
PCA+scale	0.034	0.029	0.045	0.041	0.03	0.083	0.037
PCA+dep	0.045	0.039	0.097	0.024	0.033	0.083	0.05
PCA+HDR	0.05	0.043	0.123	0.044	0.034	0.089	0.054
PCA+Cop	0.032	0.028	0.09	0.044	0.032	0.082	0.028

Table 4: PCE values for the **scm20d** dataset.

Method	Marg	Loc	Scale	Dep	PCA	HDR	Cop
None	0.167	0.104	0.355	0.036	0.093	0.405	0.245
Marg	0.154	0.117	0.34	0.027	0.088	0.379	0.245
Loc	0.152	0.074	0.336	0.026	0.086	0.359	0.236
Scale	0.156	0.109	0.31	0.027	0.084	0.365	0.246
Dep	0.167	0.116	0.346	0.026	0.094	0.4	0.245
PCA	0.163	0.113	0.344	0.027	0.092	0.387	0.245
HDR	0.159	0.109	0.34	0.025	0.09	0.375	0.244
Cop	0.162	0.108	0.339	0.033	0.09	0.381	0.243
Marg+loc	0.151	0.098	0.338	0.027	0.085	0.373	0.24
Marg+scale	0.135	0.129	0.291	0.034	0.082	0.339	0.247
Marg+dep	0.145	0.128	0.332	0.028	0.088	0.365	0.244
Marg+HDR	0.14	0.137	0.318	0.027	0.087	0.349	0.243
Marg+Cop	0.15	0.133	0.34	0.026	0.092	0.382	0.245
PCA+loc	0.155	0.088	0.33	0.022	0.088	0.366	0.237
PCA+scale	0.146	0.103	0.294	0.031	0.08	0.334	0.246
PCA+dep	0.161	0.106	0.335	0.027	0.086	0.384	0.245
PCA+HDR	0.146	0.109	0.319	0.028	0.081	0.35	0.243
PCA+Cop	0.161	0.109	0.336	0.032	0.088	0.386	0.244

Table 5: PCE values for the **wq** dataset.

Method	Marg	Loc	Scale	Dep	PCA	HDR	Cop
None	0.099	0.093	0.198	0.060	0.077	0.081	0.158
Marg	0.041	0.043	0.114	0.046	0.045	0.037	0.045
Loc	0.059	0.027	0.158	0.061	0.054	0.045	0.091
Scale	0.114	0.129	0.024	0.069	0.075	0.136	0.14
Dep	0.094	0.091	0.2	0.027	0.072	0.068	0.142
PCA	0.047	0.033	0.074	0.06	0.039	0.065	0.057
HDR	0.092	0.091	0.164	0.063	0.063	0.033	0.137
Cop	0.062	0.098	0.2	0.057	0.069	0.067	0.033
Marg+loc	0.039	0.026	0.117	0.06	0.043	0.044	0.071
Marg+scale	0.037	0.043	0.024	0.05	0.039	0.071	0.059
Marg+dep	0.04	0.043	0.105	0.024	0.043	0.042	0.049
Marg+HDR	0.042	0.047	0.124	0.046	0.046	0.039	0.049
Marg+Cop	0.041	0.052	0.129	0.044	0.049	0.036	0.032
PCA+loc	0.051	0.026	0.09	0.046	0.046	0.05	0.08
PCA+scale	0.053	0.046	0.024	0.039	0.046	0.087	0.075
PCA+dep	0.052	0.037	0.091	0.025	0.047	0.047	0.079
PCA+HDR	0.049	0.038	0.109	0.049	0.046	0.036	0.069
PCA+Cop	0.043	0.038	0.09	0.049	0.044	0.037	0.037

Table 6: PCE values for the **air** dataset.

Method	Marg	Loc	Scale	Dep	PCA	HDR	Cop
None	0.193	0.050	0.068	0.071	0.123	0.418	0.204
Marg	0.05	0.034	0.062	0.037	0.072	0.345	0.078
Loc	0.18	0.03	0.051	0.061	0.117	0.344	0.146
Scale	0.192	0.039	0.047	0.073	0.117	0.352	0.187
Dep	0.175	0.061	0.079	0.031	0.129	0.388	0.178
PCA	0.093	0.033	0.05	0.053	0.062	0.327	0.083
HDR	0.21	0.054	0.068	0.076	0.145	0.417	0.248
Cop	0.174	0.037	0.056	0.077	0.124	0.368	0.043
Marg+loc	0.033	0.03	0.061	0.047	0.052	0.355	0.047
Marg+scale	0.033	0.03	0.044	0.043	0.041	0.335	0.048
Marg+dep	0.032	0.036	0.06	0.03	0.053	0.347	0.057
Marg+HDR	0.043	0.05	0.097	0.036	0.044	0.04	0.054
Marg+Cop	0.035	0.032	0.05	0.045	0.04	0.32	0.036
PCA+loc	0.123	0.028	0.047	0.053	0.071	0.363	0.067
PCA+scale	0.118	0.031	0.043	0.057	0.067	0.371	0.065
PCA+dep	0.12	0.039	0.051	0.03	0.076	0.375	0.095
PCA+HDR	0.085	0.044	0.107	0.033	0.068	0.048	0.052
PCA+Cop	0.1	0.03	0.054	0.063	0.074	0.344	0.035

Table 7: PCE values for the **births2** dataset.

Method	Marg	Loc	Scale	Dep	PCA	HDR	Cop
None	0.095	0.092	0.125	0.022	0.066	0.097	0.149
Marg	0.028	0.029	0.057	0.019	0.029	0.032	0.028
Loc	0.05	0.023	0.08	0.021	0.039	0.039	0.088
Scale	0.102	0.131	0.021	0.028	0.067	0.073	0.118
Dep	0.098	0.097	0.127	0.019	0.067	0.073	0.154
PCA	0.049	0.025	0.043	0.027	0.025	0.038	0.073
HDR	0.097	0.102	0.076	0.025	0.059	0.029	0.141
Cop	0.046	0.062	0.124	0.018	0.055	0.066	0.031
Marg+loc	0.029	0.017	0.048	0.024	0.03	0.033	0.054
Marg+scale	0.028	0.034	0.021	0.034	0.026	0.038	0.032
Marg+dep	0.029	0.03	0.062	0.02	0.031	0.031	0.029
Marg+HDR	0.032	0.029	0.061	0.024	0.033	0.034	0.04
Marg+Cop	0.029	0.035	0.049	0.025	0.033	0.027	0.028
PCA+loc	0.043	0.02	0.062	0.031	0.032	0.033	0.067
PCA+scale	0.043	0.029	0.021	0.043	0.031	0.031	0.046
PCA+dep	0.045	0.026	0.074	0.02	0.034	0.039	0.069
PCA+HDR	0.043	0.025	0.059	0.023	0.031	0.035	0.059
PCA+Cop	0.037	0.025	0.032	0.022	0.029	0.036	0.03

Table 8: PCE values for the **households** dataset.

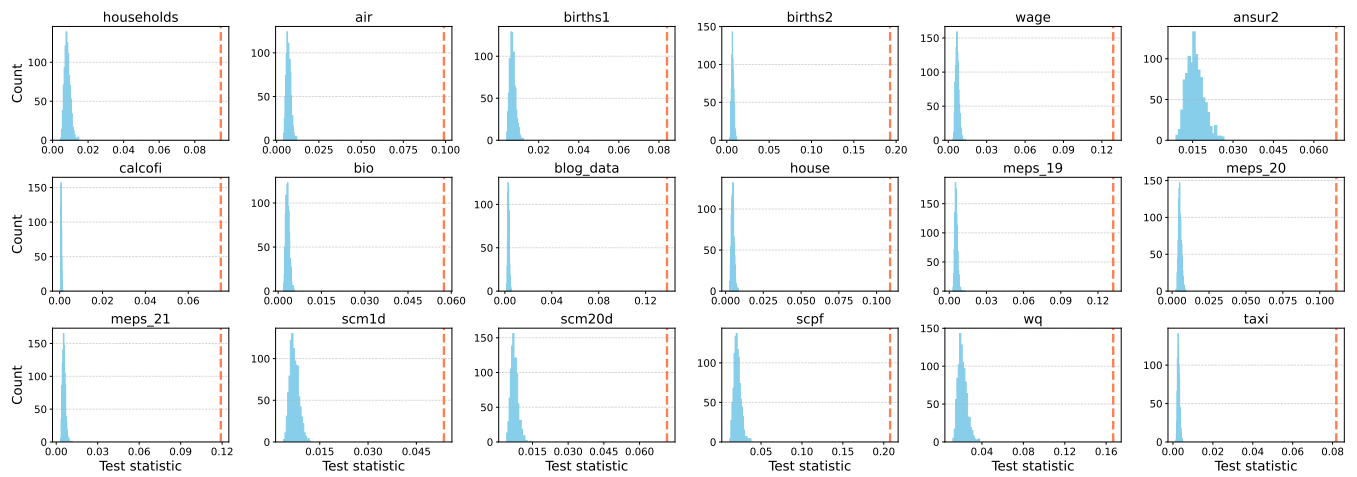


Figure 6: Distributions of the average PCE under the hypothesis of perfect calibration for all datasets, evaluated using the MIX NLL model and the **marginal** prerank.

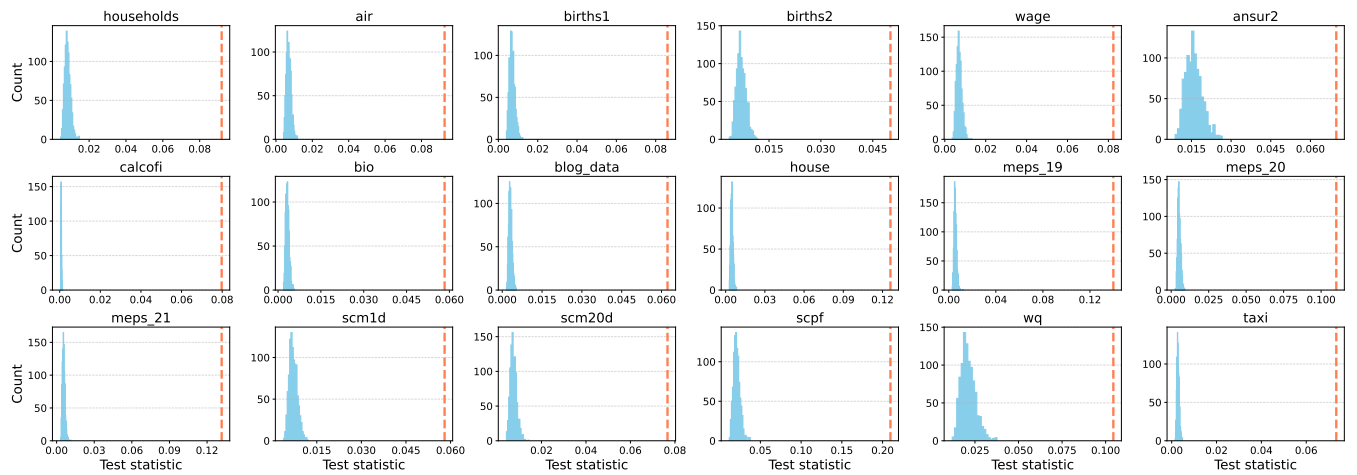


Figure 7: Distributions of the average PCE under the hypothesis of perfect calibration for all datasets, evaluated using the MIX NLL model and the **location** prerank.

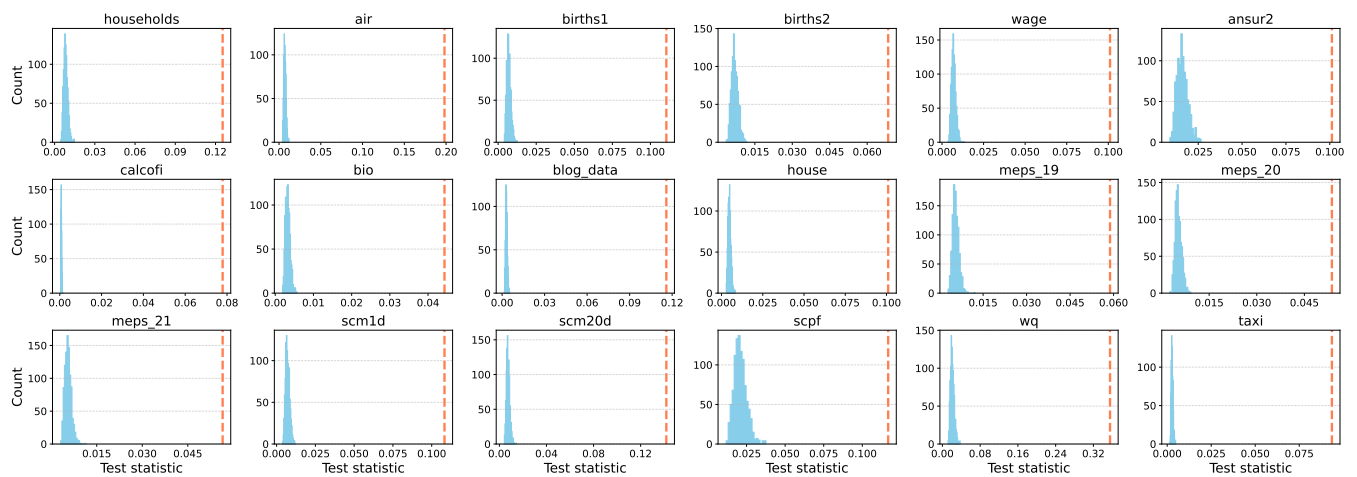


Figure 8: Distributions of the average PCE under the hypothesis of perfect calibration for all datasets, evaluated using the MIX NLL model and the **scale** prerank.

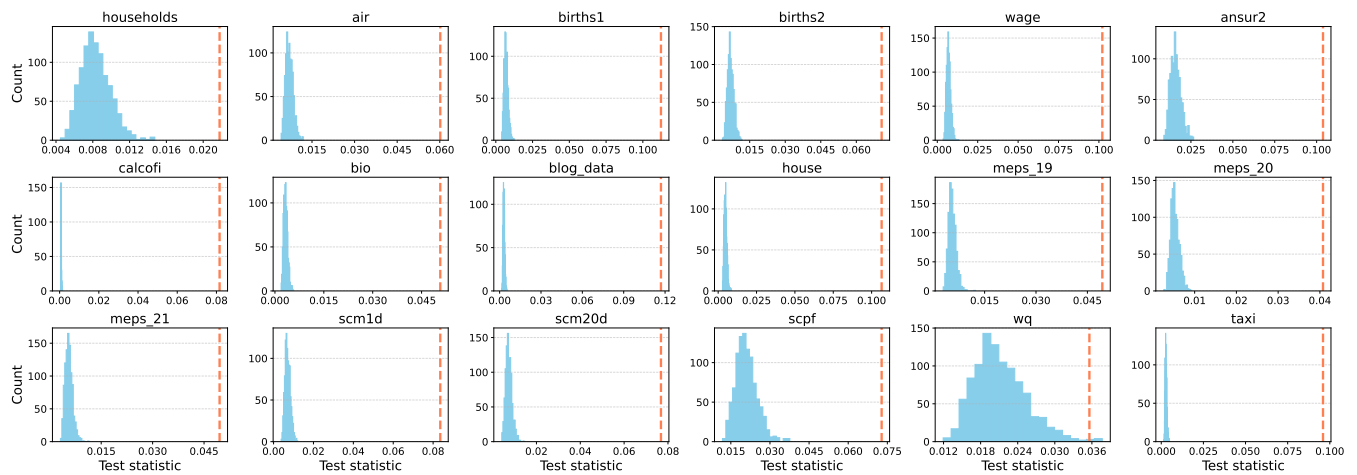


Figure 9: Distributions of the average PCE under the hypothesis of perfect calibration for all datasets, evaluated using the MIX NLL model and the **dependency** prerank.

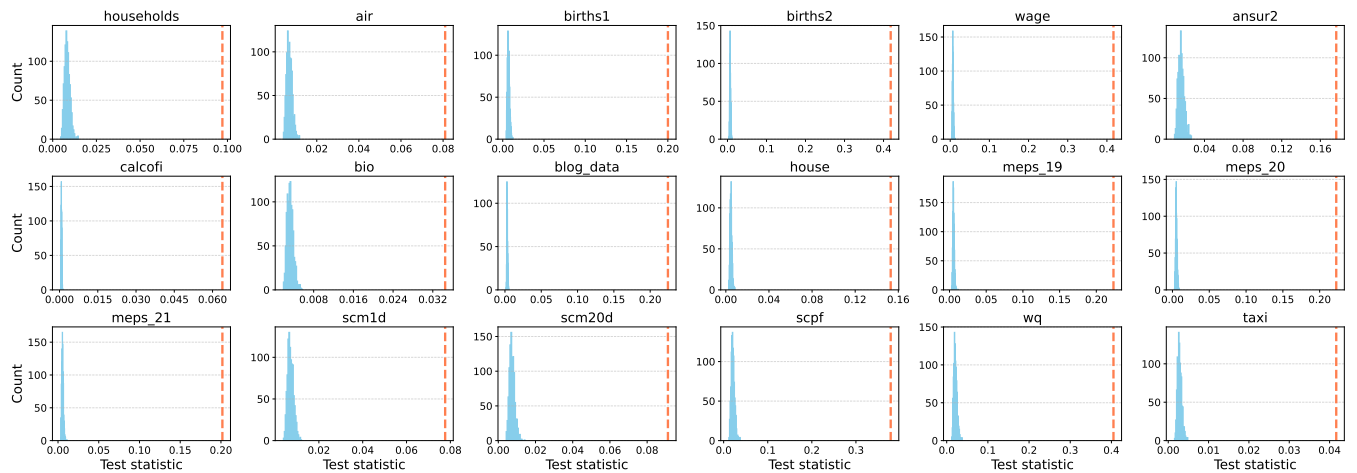


Figure 10: Distributions of the average PCE under the hypothesis of perfect calibration for all datasets, evaluated using the MIX NLL model and the **HDR** prerank.

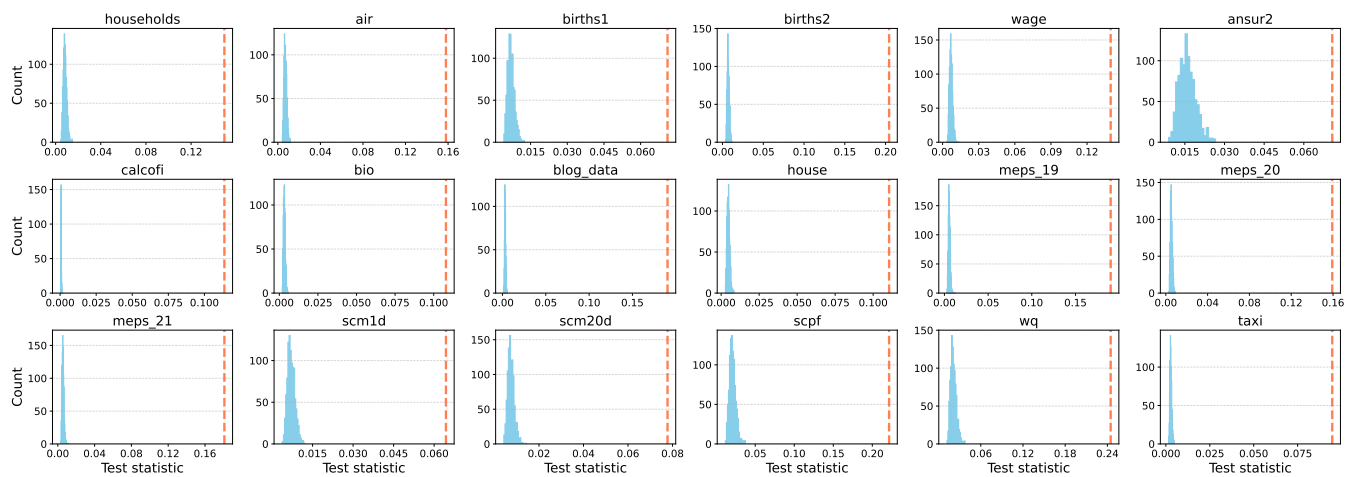


Figure 11: Distributions of the average PCE under the hypothesis of perfect calibration for all datasets, evaluated using the MIX NLL model and the **Copula** prerank.

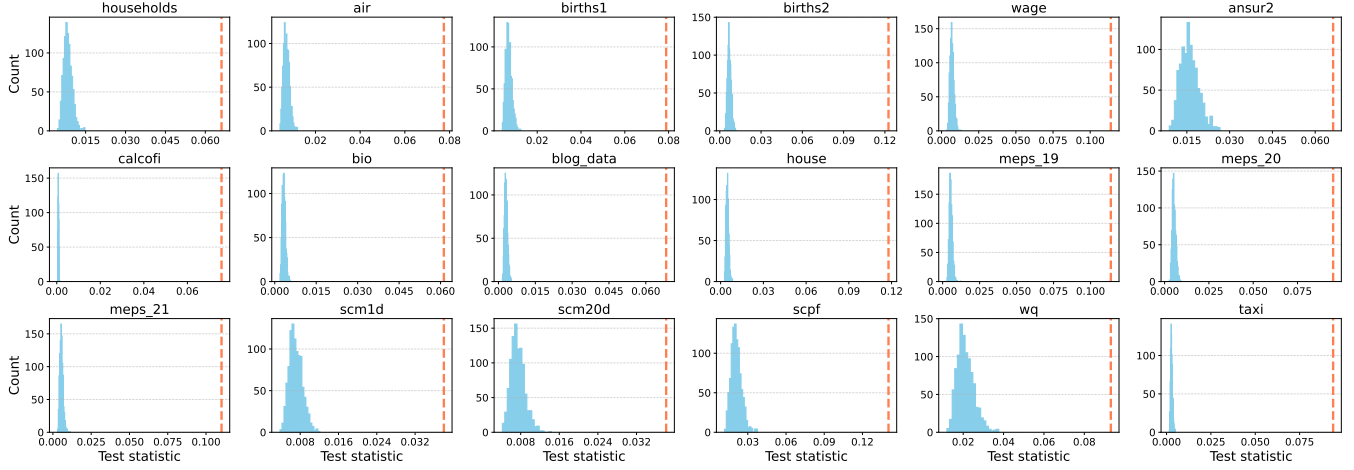


Figure 12: Distributions of the average PCE under the hypothesis of perfect calibration for all datasets, evaluated using the MIX NLL model and the **PCA** pre-rank.

Datasets	Marg.	Loc.	Scale	Dep.	PCA	HDR	Copula
households	0.095 (0.004)	0.092 (0.002)	0.125 (0.006)	0.022 (0.002)	0.066 (0.002)	0.097 (0.007)	0.149 (0.011)
air	0.099 (0.002)	0.093 (0.003)	0.198 (0.005)	0.060 (0.003)	0.077 (0.001)	0.081 (0.007)	0.158 (0.004)
births1	0.084 (0.003)	0.086 (0.003)	0.110 (0.006)	0.112 (0.006)	0.079 (0.002)	0.200 (0.011)	0.071 (0.003)
births2	0.193 (0.008)	0.050 (0.004)	0.068 (0.002)	0.071 (0.002)	0.123 (0.006)	0.418 (0.018)	0.204 (0.031)
wage	0.129 (0.001)	0.082 (0.002)	0.101 (0.004)	0.102 (0.004)	0.114 (0.003)	0.417 (0.005)	0.140 (0.003)
scm20d	0.072 (0.001)	0.077 (0.002)	0.142 (0.004)	0.077 (0.007)	0.038 (0.001)	0.091 (0.003)	0.078 (0.001)
scm1d	0.054 (0.003)	0.058 (0.005)	0.108 (0.002)	0.084 (0.009)	0.038 (0.001)	0.078 (0.004)	0.064 (0.004)
wq	0.167 (0.005)	0.104 (0.003)	0.355 (0.008)	0.036 (0.001)	0.093 (0.001)	0.405 (0.007)	0.245 (0.001)
scpf	0.208 (0.009)	0.210 (0.013)	0.117 (0.003)	0.073 (0.006)	0.140 (0.005)	0.379 (0.015)	0.222 (0.019)
meps21	0.119 (0.002)	0.132 (0.002)	0.056 (0.003)	0.050 (0.004)	0.110 (0.005)	0.202 (0.006)	0.181 (0.006)
meps19	0.132 (0.004)	0.140 (0.005)	0.059 (0.002)	0.049 (0.002)	0.114 (0.004)	0.223 (0.009)	0.190 (0.006)
meps20	0.111 (0.004)	0.110 (0.007)	0.054 (0.001)	0.041 (0.001)	0.095 (0.004)	0.223 (0.007)	0.159 (0.005)
house	0.109 (0.002)	0.126 (0.002)	0.101 (0.001)	0.107 (0.001)	0.118 (0.002)	0.153 (0.003)	0.110 (0.002)
bio	0.057 (0.002)	0.058 (0.005)	0.044 (0.002)	0.051 (0.002)	0.061 (0.003)	0.034 (0.002)	0.108 (0.005)
blogdata	0.138 (0.002)	0.062 (0.002)	0.116 (0.004)	0.117 (0.004)	0.068 (0.001)	0.224 (0.003)	0.191 (0.006)
calcofi	0.075 (0.000)	0.080 (0.001)	0.078 (0.000)	0.082 (0.000)	0.076 (0.000)	0.064 (0.001)	0.114 (0.001)
ansur2	0.068 (0.004)	0.070 (0.006)	0.101 (0.006)	0.104 (0.006)	0.066 (0.004)	0.176 (0.011)	0.071 (0.005)
taxi	0.082 (0.001)	0.073 (0.001)	0.094 (0.003)	0.096 (0.003)	0.094 (0.002)	0.042 (0.001)	0.095 (0.002)

Table 9: Results of real-world experiments using the MIX-NLL model. PCE values are computed using seven pre-rank functions across 18 real datasets and averaged over five runs. Standard errors are shown in parentheses.

Datasets	Marg.	Loc.	Scale	Dep.	PCA	HDR	Copula
households	0.030 (0.001)	0.024 (0.002)	0.021 (0.002)	0.021 (0.003)	0.026 (0.001)	0.039 (0.003)	0.031 (0.001)
air	0.040 (0.001)	0.026 (0.002)	0.027 (0.001)	0.027 (0.002)	0.039 (0.002)	0.045 (0.005)	0.029 (0.001)
births1	0.028 (0.001)	0.027 (0.002)	0.031 (0.002)	0.027 (0.002)	0.034 (0.000)	0.034 (0.003)	0.027 (0.001)
births2	0.031 (0.002)	0.029 (0.002)	0.045 (0.002)	0.032 (0.001)	0.052 (0.006)	0.428 (0.002)	0.033 (0.002)
wage	0.051 (0.020)	0.044 (0.013)	0.025 (0.001)	0.024 (0.002)	0.052 (0.008)	0.364 (0.012)	0.095 (0.015)
scm20d	0.033 (0.002)	0.025 (0.001)	0.038 (0.003)	0.022 (0.002)	0.033 (0.001)	0.093 (0.003)	0.029 (0.001)
scm1d	0.025 (0.001)	0.024 (0.002)	0.036 (0.003)	0.020 (0.002)	0.035 (0.001)	0.090 (0.015)	0.036 (0.007)
wq	0.154 (0.007)	0.076 (0.010)	0.311 (0.018)	0.028 (0.003)	0.091 (0.004)	0.376 (0.026)	0.244 (0.001)
scpf	0.039 (0.002)	0.026 (0.003)	0.041 (0.005)	0.081 (0.005)	0.063 (0.004)	0.299 (0.010)	0.032 (0.006)
meps21	0.026 (0.001)	0.025 (0.001)	0.031 (0.001)	0.024 (0.001)	0.025 (0.001)	0.032 (0.001)	0.023 (0.001)
meps19	0.026 (0.002)	0.023 (0.001)	0.050 (0.002)	0.025 (0.002)	0.024 (0.001)	0.031 (0.001)	0.022 (0.001)
meps20	0.024 (0.001)	0.023 (0.001)	0.049 (0.001)	0.033 (0.001)	0.024 (0.000)	0.034 (0.005)	0.024 (0.001)
house	0.027 (0.003)	0.020 (0.001)	0.025 (0.003)	0.020 (0.001)	0.033 (0.004)	0.028 (0.002)	0.021 (0.000)
bio	0.021 (0.001)	0.020 (0.001)	0.021 (0.001)	0.021 (0.001)	0.021 (0.001)	0.021 (0.001)	0.021 (0.001)
blogdata	0.023 (0.001)	0.024 (0.000)	0.023 (0.001)	0.023 (0.001)	0.025 (0.000)	0.030 (0.001)	0.024 (0.001)
calcofi	0.020 (0.000)	0.021 (0.000)	0.020 (0.000)	0.020 (0.000)	0.021 (0.000)	0.020 (0.000)	0.020 (0.000)
ansur2	0.032 (0.004)	0.040 (0.009)	0.031 (0.004)	0.025 (0.003)	0.038 (0.003)	0.040 (0.015)	0.039 (0.009)
taxi	0.021 (0.001)	0.022 (0.001)	0.025 (0.000)	0.023 (0.001)	0.022 (0.001)	0.022 (0.000)	0.022 (0.001)

Table 10: PCE values after applying PCE-KDE regularization with MIX-NLL using the optimal λ . Results are reported for seven pre-rank functions across 18 real datasets, averaged over five runs. Standard errors are shown in parentheses.

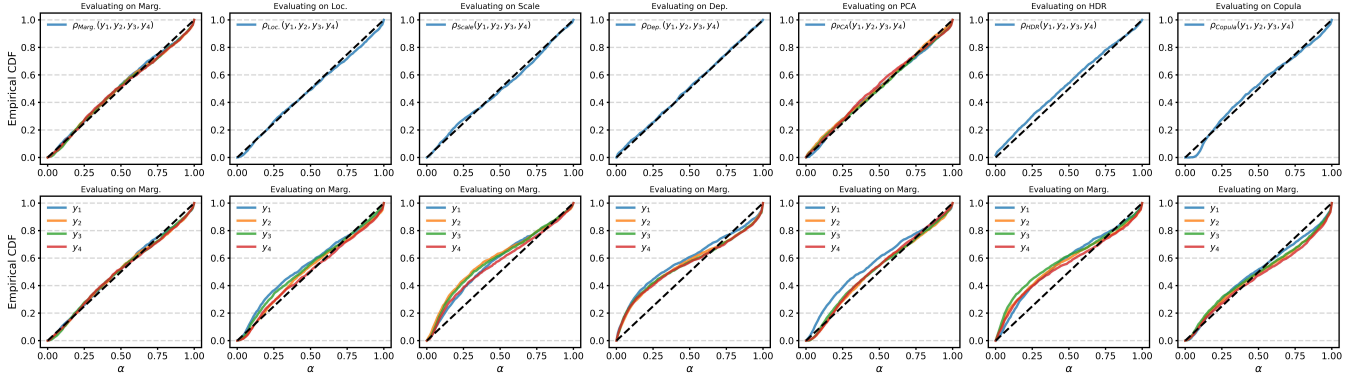


Figure 13: Reliability plots on households dataset using MIX-NLL+PCE-KDE on pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

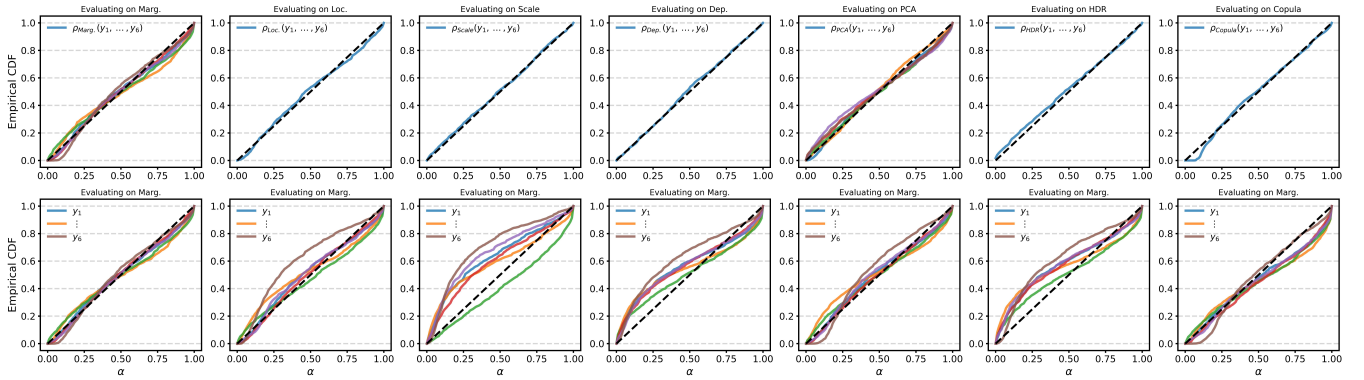


Figure 14: Reliability plots on air dataset using MIX-NLL+PCE-KDE on pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

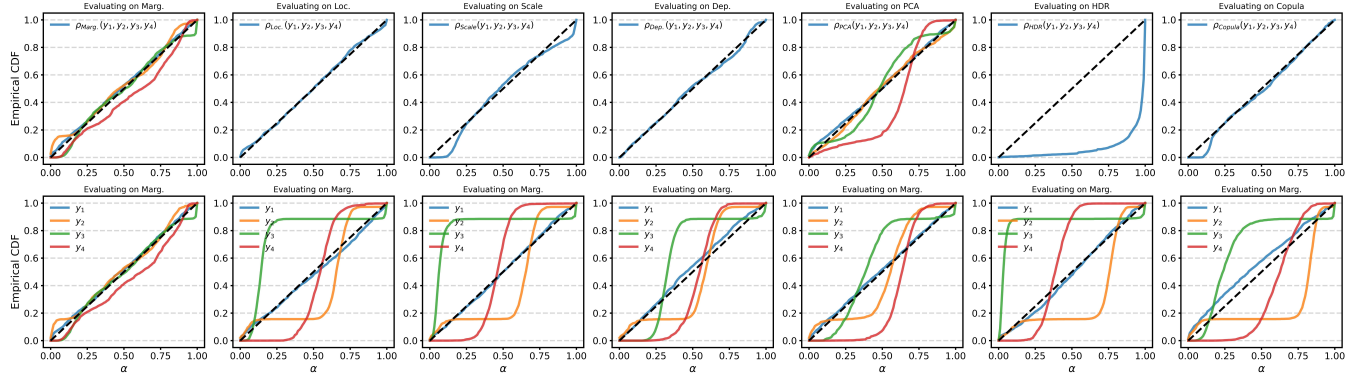


Figure 15: Reliability plots on `births2` dataset using MIX-NLL+PCE-KDE on pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

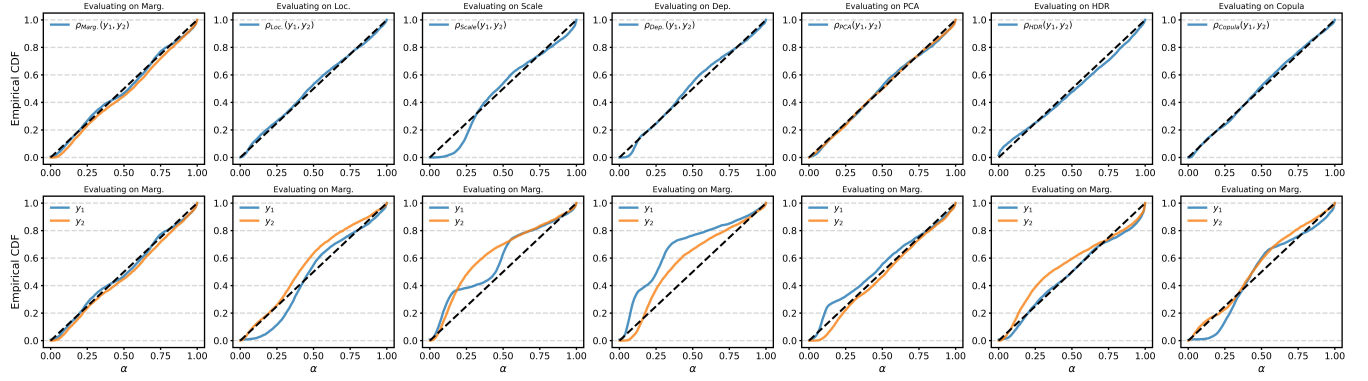


Figure 16: Reliability plots on `meps19` dataset using MIX-NLL+PCE-KDE on pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

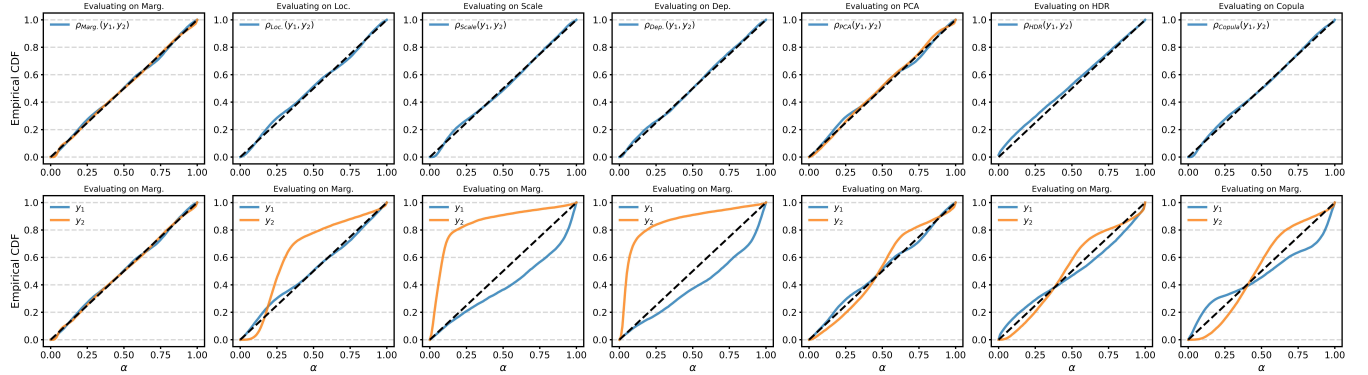


Figure 17: Reliability plots on `blog data` dataset using MIX-NLL+PCE-KDE on pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

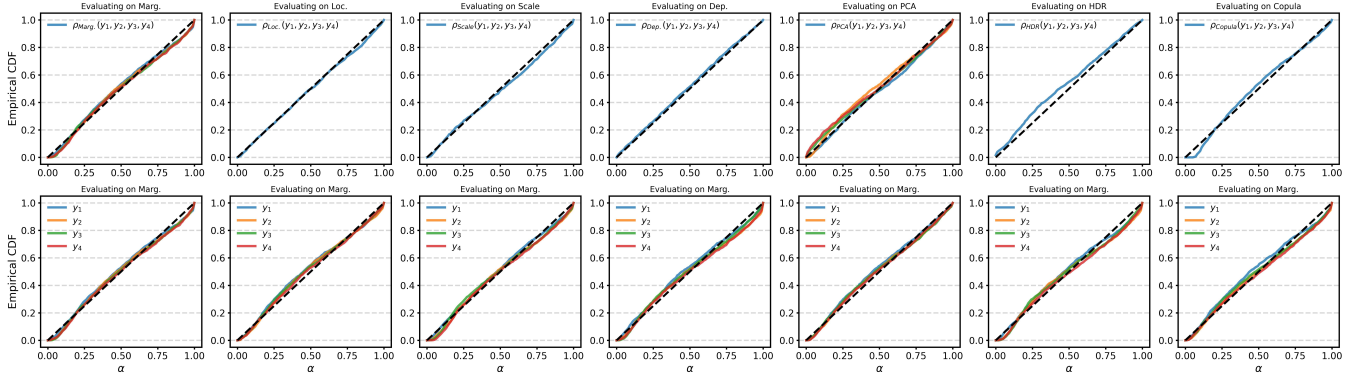


Figure 18: Reliability plots on `households` dataset using MIX-NLL+PCE-KDE on marginal+pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

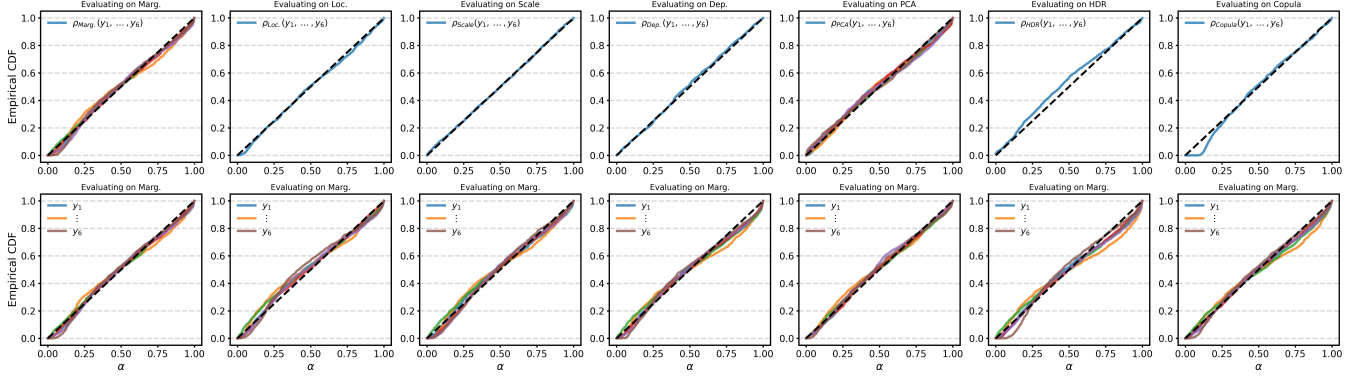


Figure 19: Reliability plots on `air` dataset using MIX-NLL+PCE-KDE on marginal+pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

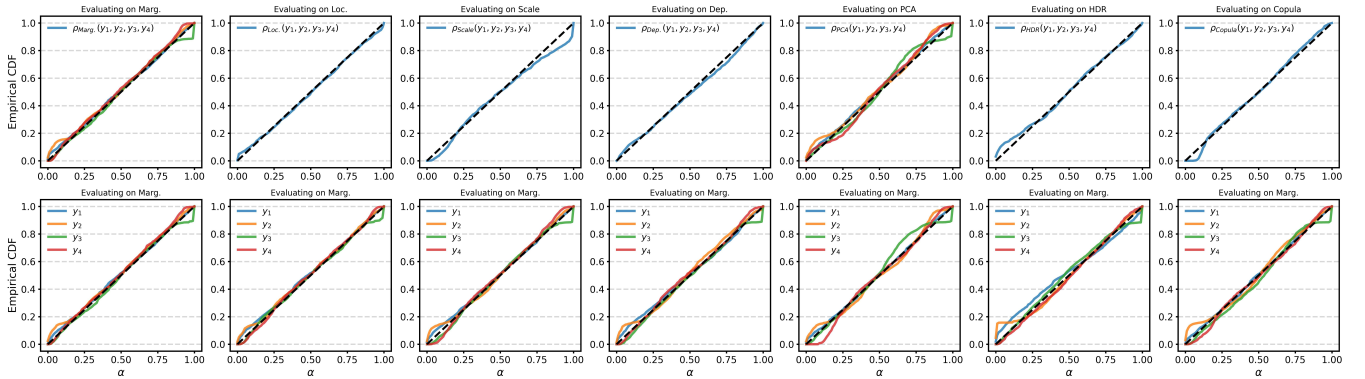


Figure 20: Reliability plots on `births2` dataset using MIX-NLL+PCE-KDE on marginal+pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

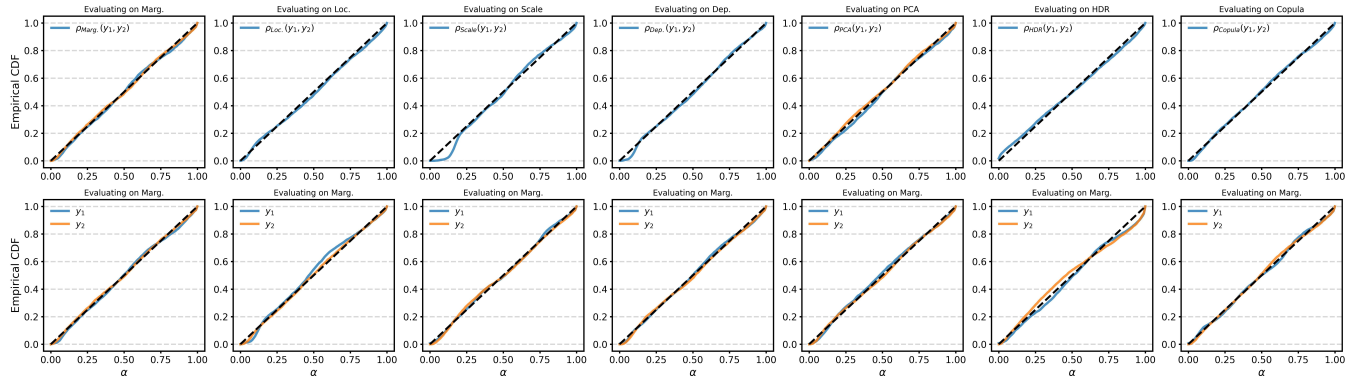


Figure 21: Reliability plots on `meps19` dataset using MIX-NLL+PCE-KDE on marginal+pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.

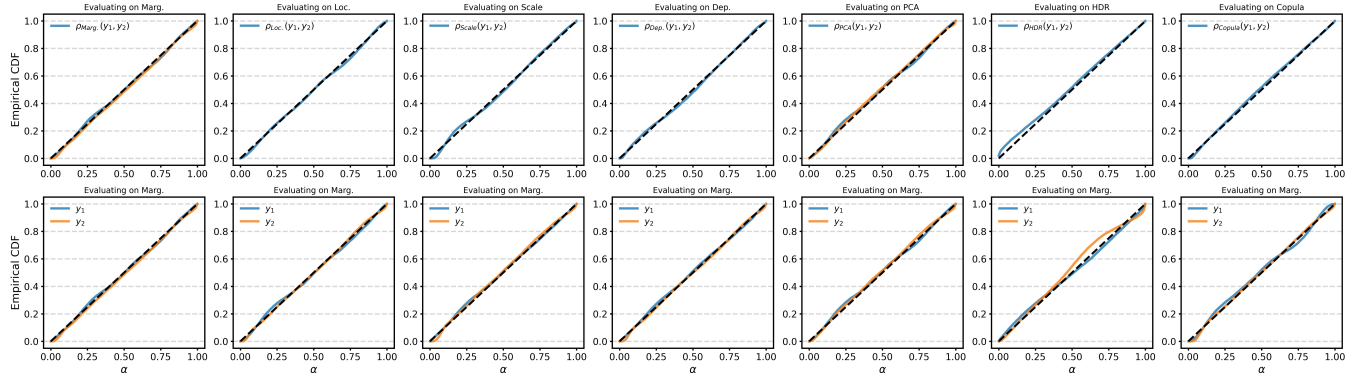


Figure 22: Reliability plots on `blog data` dataset using MIX-NLL+PCE-KDE on marginal+pre-rank. Top row: calibration curves with respect to: marginal, location, scale, dependency, PCA, HDR, and Copula. Bottom row: corresponding marginal calibration curves.